Crossmodal Sound Retrieval based on Specific Target Co-occurrence Denoted with Weak Labels

Masahiro Yasuda, Yasunori Ohishi, Yuma Koizumi, and Noboru Harada

NTT Corporation, Japan

{masahiro.yasuda, yasunori.ohishi, koizumi.yuma, noboru}@ieee.org

Abstract

Recent advancements in representation learning enable crossmodal retrieval by modeling an audio-visual co-occurrence in a single aspect, such as physical and linguistic. Unfortunately, in real-world media data, since co-occurrences in various aspects are complexly mixed, it is difficult to distinguish a specific target co-occurrence from many other non-target co-occurrences, resulting in failure in crossmodal retrieval. To overcome this problem, we propose a triplet-loss-based representation learning method that incorporates an awareness mechanism. We adopt weakly-supervised event detection, which provides a constraint in representation learning so that our method can "be aware" of a specific target audio-visual co-occurrence and discriminate it from other non-target co-occurrences. We evaluated the performance of our method by applying it to a sound effect retrieval task using recorded TV broadcast data. In the task, a sound effect appropriate for a given video input should be retrieved. We then conducted objective and subjective evaluations, the results indicating that the proposed method produces significantly better associations of sound and visual effects than baselines with no awareness mechanism.

Index Terms: crossmodal retrieval, audio-visual co-occurrence, multi-task learning, deep neural network

1. Introduction

Visual and sound events often tend to occur simultaneously: upper and lower lips move while talking; car passes on a street with engine sound; visual and sound effects are aligned in a movie. Such *co-occurrence* between vision and sound plays an important role in the way humans learn to associate visual objects to abstract concepts [1, 2]. There have been many studies on crossmodal learning using audio-visual co-occurrence. The applications of their models are crossmodal retrieval [3–12], sound source separation and localization [13–18], and audio-visual scene analysis [19–22]. In this paper, we tackle sound-effect retrieval from a video; automatically retrieving suitable sound effects corresponding to a given video input. Specifically, we use real TV broadcasting as a source of training data, which were weakly-labeled for our training purposes.

Previous studies on crossmodal learning focused on learning the co-occurrence in a single aspect, such as linguistic information [3–10] or physical information [11, 12]. Harwath *et al.* proposed an embedding model for associating visual objects with spoken words, where many pairs comprising a static image and a spoken audio caption were successfully used for a crossmodal retrieval task [3]. Owens *et al.* proposed a model to predict sound from videos as a way to study physical interactions within a visual scene, where they used hundreds of videos of people hitting, scratching, and prodding objects with a drumstick [11].

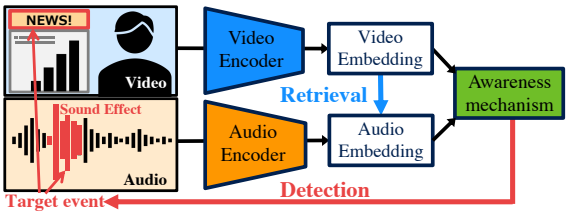

Figure 1: *System overview applying proposed method*

As a new crossmodal retrieval task, we took up real TV broadcasting as a source of learning data. In TV programs, sound effects tend to appear with other sounds, such as speech and music; therefore, the audio-visual co-occurrences in various aspects are complexly mixed. The robustness of the model with respect to noisy backgrounds is thus crucial. To model a specific audio/visual co-occurrence, we first need to identify which part of the audio/visuals should be attributed as the target. This is similar to what voice activity detection does in automatic speech recognition [23, 24] in terms of motivation. This requirement suggests that the training of a crossmodal retrieval system requires not only merely focusing on the representation learning of the target co-occurrence but also retrieving the target audio/visual event from the mixture of several co-occurrences. Practically, the system should be trained with weakly-labeled data because it is costly to annotate massive audio/visual data with strong labels.

We propose a triplet-loss-based representation learning method that incorporates an awareness mechanism. Figure 1 shows an overview of a crossmodal retrieval system with the proposed method applied. Audio and video representations are extracted with an audio encoder and video encoder respectively. The system then retrieves a suitable sound effect whose audio embedding is similar to the query video's embedding. These embeddings need to be able to distinguish between the target co-occurrence and the others. To train such robust encoders, our method incorporates an awareness mechanism, which is an additional component to detect whether the input video/audio contains the weakly-labeled target event.

2. Related Works

Recent crossmodal retrieval methods are mainly based on advancements in deep neural network (DNN)–based representation learning. In representation learning, DNNs are trained via loss function that associates between audio and visual based on similarity such as triplet loss. The simplest form of a triplet loss function can be written as:

$$\mathcal{T}_\delta(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{N}) = \max(\mathcal{D}(\boldsymbol{A}, \boldsymbol{P}) - \mathcal{D}(\boldsymbol{A}, \boldsymbol{N}) + \delta, 0), \quad (1)$$

where, $\mathcal{D}(\boldsymbol{a}, \boldsymbol{b})$ means the distance between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ defined as $\mathcal{D}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_2$, where, $\|\cdot\|_2$ is ℓ_2 norm. Three

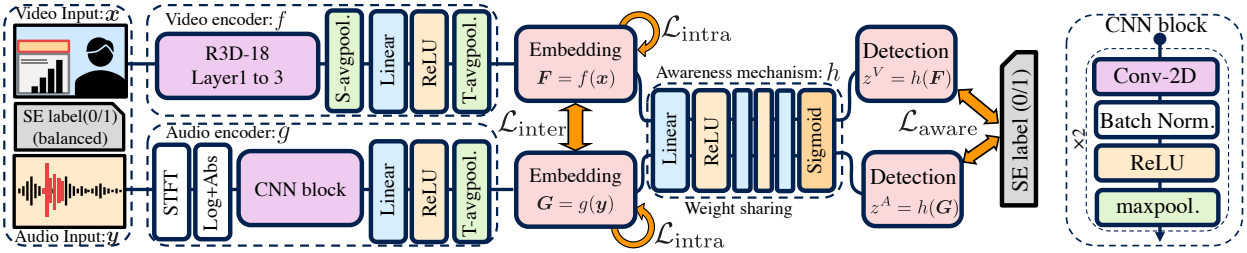

Figure 2: *Network architecture of the proposed method. "SE label" denotes binary label of sound effect."S- and T-avgpool" denotes average pooling to spatial and temporal dimensions, respectively. In CNN block part, "Conv-2D", "Batch Norm.", "maxpool" denotes 2-D convolutional layer, batch normalization, and max pooling, respectively.*

vectors $\boldsymbol{A}$, $\boldsymbol{P}$ and $\boldsymbol{N}$ are embeddings of an anchor, positive and negative samples respectively. The parameter of $\mathcal{T}$ denoted as δ is a positive constant called the margin parameter.

Linguistic-oriented crossmodal retrieval is a hot topic in this area [3–10]. Harwath *et al.* proposed a crossmodal retrieval method for direct and bidirectional retrieval between images and spoken words on an image-captioning dataset [3]. By using spoken words instead of text, it is expected that richer information can be used for retrieval such as voice pitch and speed. Not only image-audio retrieval, a recent study of linguistic-oriented crossmodal retrieval attempted to video-audio retrieval. Boggust *et al*. proposed a joint multi-modal embeddings in cooking show videos where the audio and visual streams are loosely synchronized [5].

Physical interactions between a visual scene and the corresponding sound should also be an important aspect for associating audio and visuals [11, 12]. The sound generated by the physical interactions is mainly determined by the material of the object and its action. Owens *et al.* focused on this co-occurrence and proposed a method of predicting and retrieving visually indicated sounds [11]; they modeled the physical co-occurrence between sound and image using a dataset of people hitting, scratching, and prodding various objects with a drumstick.

3. Proposed method

Our ultimate goal is to establish a crossmodal retrieval system using real-world video data in which various co-occurrence are complexly mixed. In this study, we first tackled the task of retrieving sound effects from TV broadcasting videos; automatically retrieving a suitable sound effect corresponding to a given video input.

3.1. Basic Idea

In Japanese TV programs, short sound effects of about 1 to 5 seconds are often added synchronously with various movements of video, such as, visual effects (*e.g.* appearance of superimposed captions), scene changes, and human motions. In this study, we focused on the co-occurrences across sound and visual effects. As a characteristics in sound/visual effects in TV programs, these effects are shorter than other sounds such as speech and background music (BGM), resulting in a tendency to be buried in other sounds. Therefore, our target co-occurrence exists in a mixture with other non-target co-occurrence, and we need to identify which parts of the sounds/visuals correspond to our target. Even in such a mixture of co-occurrences, humans can find and associate sound effects and visual effects. This is because humans can be aware of, and pay attention to the presence of these effects. To imitate these human abilities, we combine an awareness mechanism into conventional co-occurrence-based representation learning.

As an implementation of the above idea, we propose a multi-task learning of the audio-visual representation learning and weakly-supervised sound effect detection via an awareness mechanism. Figures 1 and 2 give an overview and the detail of the network architecture used in the proposed method. Our system consists of three parts of DNNs, i.e., video encoder, audio encoder, and the awareness mechanism. The video encoder and audio encoders encode a video and an audio signal into a shared embedding space, respectively. These encoders learn the semantic associations based on the co-occurrences across the given pair of a video and an audio. Thus, the encoders are also used in the testing-phase of sound effect retrieval. In contrast, the awareness mechanism is an auxiliary network and that is only used in the training-phase. The aim of this network is to train the encoders so as to detect the target occurrence from the mixture by identifying whether the input video/audio includes sound effects or not based on multi-task learning manner.

3.2. Implementation

Network Architecture: The video encoder f is used to embed videos $\boldsymbol{x} \in \mathbb{R}^{H\times W\times P}$ into the latent space as $\boldsymbol{F} = f(\boldsymbol{x}) \in \mathbb{R}^D$, where, H, W, and P are height, width, and the number of frames of the video. Since video effects synchronized with sound effects often have characteristic movements to attract human attention, we use 3D-convolutional neural network (CNN) as the video encoder to extract not only spatial but also temporal features. We utilized the upper 3-layers of Resnet18-3D (R3D-18) [25] pre-trained with Kinetics-400 [26] which is used for a human action classification task. The video features extracted from R3D-18 are embedded in the D-dimensional latent space through a linear layer and pooling layers.

The audio encoder g is used to embed an audio signal $\boldsymbol{y} \in \mathbb{R}^T$ into the latent space $\boldsymbol{G} = g(\boldsymbol{y}) \in \mathbb{R}^D$. Here T is the number of sample points in the time-domain. In the audio encoder, at first, the audio signal is first transformed to the log-absolute value of the short-time Fourier transform (STFT) spectrogram. The STFT spectrogram is then input to the CNN block for higher-order feature extraction. The extracted feature is next embedded in the D-dimensional latent space through a linear layer and a pooling layer. The audio encoder was pre-trained via a single-modal sound effects detection task with the awareness mechanism. This pre-training allows the audio encoder to extract features focusing on sound effects from the beginning of our multi-task learning.

The awareness mechanism is implemented by linear layers and activation functions. The activation function of the output is the sigmoid function for calculating the posterior probability of whether sound effect is exist or not. In the awareness mecha-

nism, audio and video inputs are treated independently, and the detection results are backpropagated to each encoder.

Loss Functions: The overall loss function of our proposed method consists of three terms, i.e., inter-modal triplet loss $\mathcal{L}_{\text{inter}}$, intra-modal triplet loss $\mathcal{L}_{\text{intra}}$, and awareness loss $\mathcal{L}_{\text{aware}}$. The entire network is trained in an end-to-end manner using a loss function consisting of the sum of these three terms.

To train the audio and video encoders to associate audio and visual based on co-occurrence, we used inter-modal triplet loss, which is defined as:

$$\begin{aligned}\mathcal{L}_{\text{inter}} = \sum_{m=1}^{N_b} (&\mathcal{T}_\delta(\boldsymbol{F}_{a_m}, \boldsymbol{G}_{p_m}, \boldsymbol{G}_{n_m}) \\ &+ \mathcal{T}_\delta(\boldsymbol{G}_{a_m}, \boldsymbol{F}_{p_m}, \boldsymbol{F}_{n_m})),\end{aligned} \tag{2}$$

where N_b is the batch size, a_m is the m-th anchor of triplet loss, and p_m, n_m are positive and negative samples for the m-th anchor, respectively. In inter-modal triplet loss, positive and negative labels are directly obtained from the input video and audio. If time stamps of the video and audio are the same, this pair has a positive label. If not, the pair is labelled as negative.The negative samples are selected according to the following semi-hard negative condition [27, 28] as

$$d_p < d_n < d_p + \delta, \tag{3}$$

where d_p and d_n are the ℓ_2 distance between the anchor and the positive sample and the ℓ_2 distance between the anchor and the negative sample in the latent space, respectively. This inequality shows that (i) negative samples closer to the anchor than the positive samples are not used as negative samples, and (ii) a very clear negative sample called easy-negative is excluded [27, 28].

Intra-modal triplet loss is a loss function for extracting the embedding vectors that reflects the similarity within a modal and was originally proposed for retrieval between texts and images [29]. This function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{intra}} = \sum_{m=1}^{N_b} (&\mathcal{T}_\delta(\boldsymbol{F}_{a_m}, \boldsymbol{F}_{p_m}, \boldsymbol{F}_{n_m}) \\ &+ \mathcal{T}_\delta(\boldsymbol{G}_{a_m}, \boldsymbol{G}_{p_m}, \boldsymbol{G}_{n_m})).\end{aligned} \tag{4}$$

In the intra-modal triplet loss [29], the positive sample is selected as the sample closest to the anchor in the latent space. The negative sample is selected as those that satisfies the semi-hard negative condition of (3). It plays the role of a constraint that makes intra-modal sample pairs with similar features be close to each other in the latent space.

Awareness loss is a loss function that enables the awareness mechanism to detect sound effect and written as:

$$\mathcal{L}_{\text{aware}} = \sum_{m=1}^{N_b} \text{BCE}(z^A_{a_m}, t_{a_m}) + \text{BCE}(z^V_{a_m}, t_{a_m}), \tag{5}$$

where, $z^A_{a_m}$ and $z^V_{a_m}$ are the sigmoid outputs of the awareness mechanism for m-th anchor of audio and video, respectively, and t_{a_m} is a binary weak label of sound event existence that takes one if the m-th anchor contains sound effects and zero otherwise.

Finally, to balance the triplet-based losses and $\mathcal{L}_{\text{aware}}$, we introduced a technique called margin normalization for aggregating these three losses, which is written as:

$$\mathcal{L} = \left(\mathcal{L}_{\text{inter}} + \lambda_1 \mathcal{L}_{\text{intra}}\right)/\delta + \lambda_2 \mathcal{L}_{\text{aware}}, \tag{6}$$

where, λ_1 and λ_2 are the hyper-parameters. For stabilizing training, we exponentially increase δ during training progress as [4]:

$$\delta = \delta_0 \left(\frac{\delta_{max}}{\delta_0}\right)^{\gamma}, \tag{7}$$

where, δ_0 and δ_{max} are the initial and finite value of the δ, and γ is the current number of epochs divided by the maximum number of epochs. The problem with (6) is the triplet-based losses changes their value in proportion to the margin δ, while $\mathcal{L}_{\text{aware}}$ does not. Therefore, the effect of $\mathcal{L}_{\text{aware}}$ will wane as learning progress. Thus, we normalize the triplet-based losses by dividing them by δ to balance them with the other losses.

4. Experiments

4.1. Experimental Setup

Dataset: We collected a TV Sound Effect Dataset (TVSE-Dataset) that consists of 240 hours of video broadcasted by the Japan Broadcasting Corporation (a.k.a. NHK). This dataset includes 10 days' worth of NHK broadcastings that were recorded throughout the day. Thus, it includes the various categories of TV programs such as news programs, TV dramas, comedies, and documentaries. In these TV programs, appropriate sound effects for the video had been selected by professional TV program editors. First, the 240-hour recording was divided into 6.4-second short samples, and then each short sample was weakly-labeled as to whether it contained a sound effect or not. The total number of short samples in which a sound effect was included is 4725. These samples with sound effects were divided into 3 splits on $N_{\text{train}} = 3352$, $N_{\text{valid}} = 479$, and $N_{\text{test}} = 894$ samples for training, validation, and testing, respectively. For fairness, the testing set consisted of a independent one-day broadcast videos. In the training phase, randomly selected samples without sound effects were also used. In the testing phase, only samples with sound effects were used.

Hyperparameters: For the video input, the original video of 30 fps was downsampled to 5 fps and the resolution was compressed to 224×224. Thus, $H = W = 224$ and $P = 32$. The audio sampling frequency was 48 kHz. The STFT spectrogram was then computed using a 2048 points Hanning window with a 1024 points shift.

The number of hidden units in the linear layers of video and audio encoders were 256 and 512, respectively. The number of hidden units in the linear layers of the awareness mechanism were 64, 64, 16 in order from the lower to higher layer, and the threshold of awareness mechanism was fixed to 0.5. For the 2-D convolution layers in the CNN block, we used a 3×3 kernel, a zero padding of 2, and a stride of 1, and the number of the channels was 32. The embedding dimension D was set to 64.

We used ADAM optimizer [30] with a fixed learning rate of 0.001. The weights of loss function $\lambda_{1,2}$ were set to 0.1 and 1.0 respectively, and the margin hyperparmeters δ_0 and δ_{max} were set to 1.0, 10.0 respectively. The parameters of these loss functions were determined by taking into account the retrieval accuracy of the validation split. Although the training always concludes with 40 epochs, the model of an epoch that performed the best for the validation set was used for the evaluation experiments.

Retrieval procedure: In the testing phase, sound effect retrieval was carried out in the following procedure. First, all the audio samples in the testing set were encoded to the audio embeddings, and used as the sound effect dictionary. Next, a

Table 1: *Retrieval scores of objective evaluation*

	ACC_{rank}	R@5	R@50	R@100
Random	0.500	0.006	0.056	0.112
(A) Triplet 1	0.625	0.019	0.138	0.260
(B) Triplet 2	0.685	0.024	0.191	0.320
(C) Proposed	**0.716**	**0.025**	**0.206**	**0.354**

Table 2: *Accuracy of sound event detection*

	Video	audio
(C) Proposed	0.857	0.723

video sample was encoded to a video embedding, and used as a query of sound effect retrieval. Finally, the retrieval result was obtained by sorting the ℓ_2 distances between the video query and all the audio embeddings in the sound effect dictionary.

Comparison methods: To investigate the effectiveness of our method incorporating the awareness mechanism, we decided the comparison methods based on an ablation study.

(A) Triplet 1 does not use an awareness mechanism and uses only video samples containing sound effects for training.

(B) Triplet 2 does not use an awareness mechanism and uses both video samples containing and not containing sound effects for training.

(C) Proposed is the full-architecture of the proposed method.

By comparing (B) and (C), we can directly observe the changes in their performance with and without the awareness mechanism. However, without using the awareness mechanism, it is not necessary to use a sample without sound effects for learning, and this may degrade the performance of (B). Therefore, we included (A) for a fair comparison. As the worst case of the sound effect retrieval, we refer to a random method as a baseline. This is a method of ordering retrieval results in a random order.

4.2. Objective evaluation

We conducted an objective evaluation using recall@K and ranking accuracy as the metrics for objective evaluation. The recall@K is the rate that the ground-truth audio files are within the K-th rank of the retrieval result. The ground-truth of the retrieval task was set to the audio-visual pair obtained from the same sample. Since the evaluation based on Recall@K depends on the choice of K, we also used a metric independent of K, called the ranking accuracy [31]. The ranking accuracy is defined as:

$$ACC_{\text{rank}} = \frac{1}{N_{\text{test}} + 1} \sum_{N_{\text{test}}} \left(\mathcal{R}_{\text{gt}} - 1 \right), \tag{8}$$

where $\mathcal{R}_{\text{gt}}$ is the rank of the ground-truth.

Table 1 lists the results of the objective evaluation. The results show that (C) outperforms the other methods in all the metrics. From comparing with (B) and (C), we can confirm that introducing the awareness mechanism is effective in improving the performance. From comparing (A) and (B), the performance improvement was observed by using the samples without sound effects. We consider this may indicate that the difference between the samples with sound effects and the samples without sound effects were implicitly learned and the attention to sound effects was emphasized on the retrieval task.

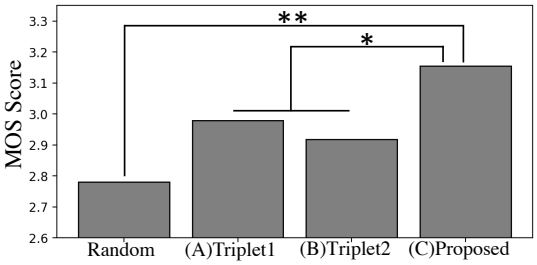

Figure 3: *Subjective evaluation results. The symbols ** and * represent $p < 0.01$ and $p < 0.05$ in the one-sided Mann–Whitney U-test, respectively.*

Although the results in Table 1 indicate the effectiveness of the awareness mechanism, they do not indicate whether the awareness mechanism component works properly. To confirm that the awareness mechanism works to find the sound effect, we evaluated the detection accuracy. Table 2 lists the accuracy in sound effect detection from audio or video of the awareness mechanism used in (C). These results show that the awareness mechanism works properly to detect sound effects.

4.3. Subjective evaluation

To verify the performance of the proposed method in terms of human sense, we conducted a subjective evaluation experiment. The evaluation samples were made by combining the sound effect with the highest retrieval rank for each video. Note that, the timing of sound effects was manually aligned between audio and video for the subjective evaluation samples. The evaluation was carried out by 18 participants. The participants watched each of the video samples generated by the proposed and comparison methods. Then they gave a five-point rating to each sample in terms of appropriateness of sound effects for the video, where each point means 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad, respectively. Using the obtained scores for each video sample, we calculated the mean-opinion-score (MOS) for each method.

Figure 3 shows the results of the subjective evaluation. The results show that the proposed method achieved the highest MOS. In one-sided Mann–Whitney U-test, statistically significant differences ($p < 0.05$) were observed between (C) and other comparison methods. Significant statistical differences were not observed between the comparison methods and the random method. These results indicate that the attention to the specific co-occurrence by introducing the awareness mechanism is effective for the retrieval of more appropriate sound effects in terms of human sense.

5. Conclusions

In this study, we proposed a triplet-loss-based representation learning method that incorporates an awareness mechanism and applied it to the sound effect retrieval task using TV broadcast data. The audio and video representations were extracted using DNN-based audio/video encoders and embedded in a shared latent space. The awareness mechanism receives these embeddings and detects the weakly-labeled sound effects. We evaluated the performance of our method by applying it to a sound effect retrieval task using recorded TV broadcast data. Objective and subjective experiments showed that the proposed method produces significantly better associations of sound and visual effects than baselines with no awareness mechanism. Therefore, we conclude that the proposed method is effective for crossmodal retrieval.

6. References

[1] E. Dupoux "Cognitive science in the era of artificial intelligence: A Roadmap for Reverseengineering the Infant Language-learner" in *Cognition, Vol. 173, p43–59*, 2018

[2] E. S. Spelke "Principles of object perception" in *Cognitive Science, Vol. 14, p29–56*, 1990

[3] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass "Joint Discovering Visual Objects and Spoken Words From Raw Sensory Input" in *Proc. of Euro. Conf. on Computer Vision (ECCV)*, 2018.

[4] G. Ilharco, Y. Zhang,and J. Baldridge "Large-scale Representation Learning From Visually Grounded Untranscribed Speech" in *Proc. of the SIGNLL Conf. on Computational Natural Lang. Learning (CoNLL)*, 2019.

[5] A. Boggust, K. Audhkhasi, D. Joshi, D. Harwath, S. Thomas, R. Feris, D. Gutfreund, Y. Zhang, A. Torralba, M. Picheny, and J. Glass "Grounding Spoken Words in Unlabeled Video" in *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] E. Azuh and D. Harwath and J. Glass "Towards Bilingual Lexicon Discovery From Visually Grounded Speech Audio" in *Proc. of Interspeech*, 2019.

[7] H. Kamper and A. Anastassiou and K. Livescu "Semantic Query-by-example Speech Search Using Visual Grounding" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019.

[8] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath and J. Glass "Trilingual Semantic Embeddings of Visually Grounded Speech with Self-attention Mechanisms" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020.

[9] D. Harwath and W. H. Hsu and J. Glass "Learning Hierarchical Discrete Linguistic Units from Visually-grounded speech" in *Proc. of the Int'l Conf. on Learning Representations (ICLR)*, 2020.

[10] Y. Ding, Y. Xu, S. Zhang, and Y. Cong "Self-Supervised Learning for Audio-Visual Speaker Diarization" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020.

[11] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman "Visually Indicated Sounds" in *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] H. Zhao, C. Gan, W. Ma, and A. Torralba "The Sound of Motions" in *Proc. of Int'l Conf. of Computer Vision (ICCV)*, 2019.

[13] A. Senocak, T. Oh, J. Kim, M. Yang, and I. S. Kweon "Learning to Localize Sound Source in Visual Scenes" in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba "The Sound of Pixels" in *Proc. of Euro. Conf. on Computer Vision (ECCV))*, 2018.

[15] "R. Gao and R. Feris and K. Grauman" "Learning to Separate Objects Sounds by Watching Unlabeled Video" in *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] R. Arandjelović and A. Zisserman "Objects that Sound" in *Proc. of the Euro. Conf. on Computer Vision (ECCV)*, 2018.

[17] A. Ephrat and I. Mosseri and O. Lang and T. Dekel and K. Wilson and A. Hassidim and W. T. Freeman and M. Rubinstein "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation" in *Proc. of the ACM Special Interest Group on Computer Graphics (SIGGRAPH)*, 2018.

[18] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, A. Torralba "Self-Supervised Audio-Visual Co-segmentation" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019.

[19] Y. Aytar, C. Vondrick, and A. Torralba "SoundNet: Learning Sound Representations from Unlabeled Video" in *Proc. of Ann. Conf. on Neural Information Process. System (NIPS)*, 2016.

[20] J. Cramer, H. Wu, J. Salamon, and J. P. Bello "Look, Listen, and Learn More: Design Choices For Deep Audio Embeddings" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019.

[21] A. Owens and A. A. Efros, "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features" in *Proc. of the Euro. Conf. on Computer Vision (ECCV)*, 2018.

[22] A. Owens and J. Wu and J. H. McDermott and W. T. Freeman and A. Torralba "Ambient Sound Provides Supervision for Visual Learning" in *Proc. of the Euro. Conf. on Computer Vision (ECCV)*, 2016.

[23] Y. Tachioka "DNN-Based Voice Activity Detection Using Auxiliary Speech Models in Noisy Environment" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2018.

[24] R. Masumura, K. Matsui, Y. Koizumi, T. Fukutomi, T. Oba and Y. Aono "Context-Aware Neural Voice Activity Detection Using Auxiliary Networks for Phoneme Recognition, Speech Enhancement and Acoustic Scene Classification" in *Proc. of the 27th Euro. Signal Process. Conf. (EUSIPCO)*, 2019.

[25] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri "A Closer Look at Spatiotemporal Convolutions for Action Recognition" in *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hiller, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman "The Kinetics Human Action Video Dataset" in *Computing Research Repository, abs/1705.06950*, 2017.

[27] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler "VSE++: Improving visual-semantic embeddings with hard negatives" in *Proc. of the British Machine Vision Conf. (BMVC)*,2018.

[28] A. Jansen, M. Plakal, R. Pandya, D. P. Ellis, S. Hershey, and J. Liu "Unsupervised Learning of Semantic Audio Representations" in *Proc. of IEEE Int'l Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2018.

[29] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao "Triplet-Based Deep Hashing Network for Cross-Modal Retrieval" in *IEEE Trans. on Image Processing Vol.27*, 2018.

[30] D. P. Kingma and J. Lei. Ba "Adam: A Method for Stochastic Optimization" in *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2015.

[31] F. F. Kuo, M. K. Shan, and S. Y. Lee "Background Music Recommendation for Video Based on Multimodal Latent Semantic Analysis" in *Proc. of IEEE Int'l Conf. on Multimedaia and Expo (ICME)*, 2013.